УДК 004.9 **doi: 10.26906/SUNZ.2024.2.163**

О. В. Юдін, М. Ю. Шипунов

Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Харків, Україна

МЕТОДИ ПРОТИДІЇ АВТОМАТИЗОВАНОМУ ЗБОРУ ІНФОРМАЦІЇ З ВЕБСТОРІНОК

Анотація. Робота присвячена аналізу методів протидії автоматизованому скануванню веб-вузлів. Метою роботи є аналіз алгоритмів та особливостей роботи парсингових систем та, на основі отриманих даних, побудова системи засобу, який буде спеціалізуватися саме на виявленні та протидії атак з використанням парсингових систем. Методом дослідження виступає аналіз методів протидії парсинговим системам. У роботі було розглянуто історію виникнення автоматизованих систем, їх класифікацію, особливості роботи та методи протидії. Запропоновані методи спеціалізуються на захисті від парсингових систем, та створює мінімальне додаткове навантаження на серверне обладнання, що не заважає роботі звичайним користувачам. Дані методи будуть корисними власникам великих ресурсів, на яких головним є саме інформація.

Ключові слова: парсингове програмне забезпечення, методи протидії ботам, парсери, crawler.

Вступ

Розвиток інформаційних технологій ставить все більш складні завдання стосовно обробки інформації, її зберігання, передавання та забезпечення захисту [1]. Одночасно з цими завданнями виникла можливість перекласти частину рутинних завдань на обчислювальну техніку. З плином часу потужності комп'ютерів зростали і розробники все частіше створювали програми для вирішення простих, тривіальних завдань. Пізніше такі програми отримали власну назву – боти. На сьогоднішній день, з розвитком якості глобальної мережі Інтернет кількість таких рутинних завдань лише збільшується. Наразі будь яка сфера так чи інакше взаємодіє з комп'ютерними системами. Наприклад на будь-якому складі встановлена система обліку об'єктів. У магазинах встановлені термінали оплати, які пов'язані з базою товарів. На великих підприємствах зазвичай працює поштовий сервер для ефективного обміну інформацією між співробітниками та керівництвом.

Для більш швидкого виконання цих завдань створюють різних ботів-помічників. Таким чином задля прискорення додавання нового товару на склад була створена система штрих-кодів, які скануються і автоматично додаються до потрібної категорії.

Наразі індустрія ботів активно розвивається. За останні роки зацікавленість до теми автоматизованих помічників лише зростає. За даними компанії IBM, станом на листопад 2023 року, 1,4 мільярди користувачів використовують ботів. У свою чергу компанії почали використовувати автоматизовані системи для організації технічної підтримки. Близько 85% питань від клієнтів наразі вирішують чат-боти [2].

Існує велика кількість ботів, як для комерційного використання, так і користувацькі. На сьогоднішній день, ChatGPT, головна розробка компанії OpenAI користується популярністю серед простих користувачів. Це зумовлено її здатністю вирішувати різноманітні завдання. Від пошуку і фільтрації інформації до генерації нового контенту (тексти, вірші або навіть коди програм). Але нажаль автоматизовані системи приносять не лише користь, а й додають нові проблеми людству. Згідно звіту компанії Imperva за 2023 рік, доля ботів у глобальній мережі складає 47,5% [3]. У свою чергу більше половини таких систем є шкідливими. Це призводить до генерації великої кількості додаткового трафіку, для маршрутизації якого необхідні додаткові потужності. Окрім цього, все більше з'являється судових позивів щодо авторських прав на контент, згенерований за допомогою штучного інтелекту (ШІ). Наприклад розробники соціальної мережі Instagram зобов'язались додати сповіщення користувачам про те, що поточний контент згенерований за допомогою нейронних мереж [4]. Окрім проблем, з якими людство вже зіштовхнулось існують і ті, які дадуть про себе знати у майбутньому.

Наприклад заміна працівників деяких сфер нейронною мережею в подальшому може призвести до появи нестандартних проблем, які система не здатна вирішувати. На сьогоднішній день, студенти та працівники IT сфери активно користуються ChatGPT для вирішення робочих завдань. В подальшому це може призвести до виявлення алгоритмічних закладок, про які ніхто не знає [5]. Окрім цього старі проекти перестануть підтримуватися, оскільки не залишиться спеціалістів, які починали створювати даний проект.

Однією з найважливіших проблем є використання ботів під час атак на веб-ресурси. Активний розвиток автоматизованих систем лише додає складностей, оскільки боти все ближче наближаються до вирішення тесту Т'юринга, що призводить до ускладнень їх виявлення. Власники інформаційних веб-ресурсів дізнаються про атаки, які були здійснені за допомогою парсинговий систем (ПС), лише після того, як помічають наявність їх корпоративної інформації на сторонніх ресурсах у даркнеті [6].

Виходячи з цього, метою даної роботи є: аналіз алгоритмів та особливостей роботи ПС та, на основі отриманих даних, побудова системи засобу, який буде спеціалізуватися саме на виявленні та протидії атак з використанням ПС.

Терміни та визначення

Чат-боти – це комп'ютерні програми, які імітують людську мову (усну чи письмову) та дозволяють спілкуватися з цифровими пристроями так, якби вони були живими людьми. Чат-боти можуть бути дуже простими, як елементарні програми, що відповідають на простий запит однорядковою відповіддю, або

складними, як цифрові помічники, які навчаються та розвиваються в міру збору та обробки інформації, тим самим підвищуючи свій рівень персоналізації [7].

Парсери – це комп'ютерні програми для автоматизованого збору та систематизації інформації, розміщеної на різних сайтах.

Індексатор – це робот, який зазвичай використовується пошуковою системою для виявлення нових сторінок в інтернеті. Принцип роботи таких ботів полягає у постійному скануванні сторінок та виявленню на них посилань з подальшим переходом по них. Всю зібрану інформацію робот заносить до спеціальної бази даних, яка називається індексом. Дані про нові сторінки в інтернеті пошукова машина бере саме з такого індексу [8].

Генератори контенту – це автоматизовані системи, які мають на меті створення нової інформації ґрунтуючись, зазвичай, на текстовому запиті. Найчастіше такі боти у своїй основі мають нейромережу, що означає, що при кожній відправці одного і того ж самого запиту буде повертатися відповідь, яка відрізняється від попередньої. [9].

Пошукові – це автоматизовані системи, які створені з метою допомогти користувачам шукати необхідний фрагмент інформації у великій кількості даних. Зазвичай такі боти працюють разом з краулерами, що дозволяє отримувати актуальні дані щодо необхідного запиту користувача. [10].

Аналіз атак, спричинених ботами

На сьогоднішній день, основним показником розвитку комп'ютерних систем є ріст їх продуктивності. Нажаль, отримані потужності не завжди використовуються з добрими намірами. Зловмисники отримали потужний інструмент для швидкої реалізації експлойтів (metasploit) та повторювання операцій (спам-розсилки). У 2004 році був створений комп'ютерний хробак Bagle, який розповсюджувався за допомогою електронної поштою. Вкладення у письмі було файлом .EXE з іконкою програми Windows Calculator. При завантаженні вкладення на пристрій хробак відкривав програму калькулятора, а також потай від користувача копіював себе у директорію під назвою «bbeagle.exe» і створював певні ключі реєстру [11]. Після цього хробак завантажував бекдор, який забезпечував віддалений доступ до зараженого комп'ютера за портом 6777. Потім хробак шукав електронні адреси у файлах адресної книги Windows і продовжував поширюватися електронною поштою. Важливо зазначити, що ігнорувалися електронні адреси, які містили у назві рядок «.r1», «@windows», «@avp», «@hotmail.com» або «@msn.com» [12]. Створена ботнет мережа використовувалася зловмисниками для розсилки спам повідомлень та виконання DDoS атак.

У 2008 році 18-річний студент Оуен Уокер створив Akbot – вірус, головною метою якого було створення ботнет мережі. Під час зараження, пристрій додавався до серверів IRC та очікував вказівок. За час свого існування мережа складалася з 1,3 мільйонів пристроїв та використовувалася для реалізації DDoS атак [13]. На початку 2012 року фахівці ESET виявили ботнет мережу, яка отримала назву Win32/Georbot. За даними компанії ESET, команди, що управляють, виходять з офіційного сайту уряду Грузії. Метою створення даного ботнета є викрадення документів та цифрових сертифікатів із заражених комп'ютерів. Ще одна особливість цієї шкідливої програми полягає у тому, що вона шукає на інфікованому пристрої файли конфігурації RDP (Remote Desktop Connection) з метою подальшого розкрадання та отримання несанкціонованого доступу до них [14]. Крім того, бот здатний створювати аудіо- та відео-записи та збирати інформацію про локальну мережу.

У 2012 році шість великих американських банків – Bank of America, JPMorgan Chase, US Bank, Citigroup, Wells Fargo та PNC Bank – стали жертвами DDoS-атаки, яка тривала кілька тижнів і заважала нормальному функціонуванню їхніх сайтів та онлайн-сервісів [15]. Атаку організували хакерська група, яка називає себе Izz ad-Din al-Qassam Cyber Fighters, яка заявила, що її мотивом був протест проти антимусульманського фільму «Невинність мусульман». Атака досягала потужності до 60 Гбіт/с. У 2016 році ботнет Mirai, який складався з мільйонів заражених пристроїв інтернету речей, таких як камери та маршрутизатори, провів кілька масштабних DDoS-атак на різні цілі [16]. Серед них були сайт фахівця з кібербезпеки Брайана Кребса, DNS-провайдер Dyn, який обслуговував багато популярних сайтів, таких як GitHub, HBO, Twitter, Reddit, PayPal, Netflix та Airbnb, а також французький хостинг-провайдер OVH. Потужність атак досягала від 623 Гбіт/с до 1.1 Тбіт/с. У 2017 році Google зазнала найпотужнішої DDoS-атаки в історії, яка досягла 2.54 Тбіт/с. Атака була частиною піврічної кампанії, яку вели урядові китайські хакери проти інфраструктури Google [17]. Однак Google змогла відбити атаку та забезпечити нормальну роботу своїх сервісів та безпеку даних користувачів. У 2018 році спеціалісти з Check Point викрили шкідливе програмне забезпечення, яке поширювалось серед Android пристроїв та об'єднувало їх у ботнет мережу. Спеціалісти зазначили, що на час знайдення, вже було заражено близько 5 мільйонів пристроїв [18]. У 2020 році у Гонконгу була зафіксована атака на системи відеоспостереження міста з використанням машинного зору. Таким чином зловмисники слідкували за переміщенням поліцейських патрулів та інформували протестувальників про їх наближення [19]. У 2020 році, згідно даних компанії Positive Technologies, було виявлено глобальну кампанію, в якій зловмисники використовували веб-краулери для збору даних про мільйони користувачів соціальних мереж, таких як Facebook, Instagram, Twitter і YouTube. Ці дані були використані для фішингових атак, маніпуляції громадською думкою, шантажу тощо [20]. У 2021 році, згідно даних компанії Positive Technologies, була зафіксована глобальна атака на веб-сайти державних установ різних країн, у тому числі США, Німеччини, Франції та ін. впроваджували шкідливий код на ці сайти. Метою атаки було зараження відвідувачів сайтів та отримання доступу до їх пристроїв [21].

Компанія StormWall 4 серпня 2021 повідомила про виявлення DDoS-атак, організованих, за словами її фахівців, за допомогою «найпотужнішого ботнету за весь час існування інтернету». Таким чином,

максимальна потужність атаки сягала 2 Тбіт/с, що станом на 4 серпня було абсолютним рекордом серед усіх атак за участю ботнетів. Більшість DDoS-атак були спрямовані на ігрову індустрію. Ботнет складався з 49 тисяч пристроїв, серед яких знаходяться лише сервери, звичайні комп'ютери і мобільні пристрої відсутні. Атаки, які здійснювалися даним ботнетом, були стандартні, було запущені атаки протоколами UDP, TCP і HTTP з емуляцією браузера [22].

У 2022 році, згідно даних компанії Positive Technologies, було виявлено глобальну атаку на вебсайти блокчейн-проектів, таких як криптовалютні біржі, гаманці, платформи і т.д [23]. Зловмисники використовували веб-краулери для збору інформації про користувачів, таких як адреси електронної пошти, паролі, номери телефонів тощо. Потім вони використовували цю інформацію для атак по ланцюжку, тобто для злому інших акаунтів і сервісів, пов'язаних з первинними жертвами. Метою атаки було крадіжка криптовалют та інших цифрових активів.

У 2022 році дослідники Fortinet FortiGuard Labs виявили шкідливу активність, яка була призведена ботнетом на основі Golang, який зламував сайти WordPress, щоб захопити контроль над цільовими системами [24]. Перша версія була здатна лише атакувати SSH порти потенційного Linux серверу. Оновлена версія мала додатковий функціонал для брутфорсу за допомогою telnet та через протокол тунелювання Generic Routing Encapsulation [25].

У квітні 2022 року фахівці ІБ-компанії Fortinet виявили ботнет на базі початкового коду Mirai, який отримав назву Enemybot. EnemyBot – це небезпечний ботнет, який кіберзлочинці використовують у першу чергу для запуску DDoS-атак [26]. Даний ботнет був націлений на широкий спектр маршрутизаторів та пристроїв Інтернету речей (IoT), які використовували застаріле програмне забезпечення або примітивні облікові дані для входу.

Як видно з наведених прикладів, використання ботів ефективно для організації атак, які впливають на головні складові інформаційної безпеки – доступність, конфіденційність і цілісність.

Класифікація атак, спричинених ботами

На сьогоднішній день, існує велика кількість вузькоспеціалізованих шахрайських ботів, створених для конкретних завдань. Це призводить до складнощів під час створення конкретної класифікації цих систем. У представленій роботі буде розглянуто наступну класифікацію атак з використанням ботів:

- широкомасштабні;
- вузько спрямовані;
- гібридні;
- з використанням машинного зору;
- з використанням ботнет мереж.

Широкомасштабні атаки – головною метою даного виду атак є зараження якомога більшої кількості пристроїв. Зазвичай для виконання цієї атаки використовуються найпростіші боти, або навіть невеликі скрипти. Прикладом широкомасштабної атаки є спам розсилка. Під час розсилки спаму, головна задача зловмисника відправити заражений лист найбільшій кількості користувачам. Завдяки цьому зловмисник збільшує шанси зустріти неосвіченого користувача, який відкриє вірусне посилання або файл.

Вузько спрямовані атаки – метою даної атаки є автоматична перевірка можливості використання певної вразливості. Таким чином, такі автоматизовані системи допомагають автоматизувати підготовку до глобальної атаки або протестувати виконання зараженого скрипту, створеного зловмисником на потенційно зараженому пристрої. Прикладом такого виду атак є експлуатація вразливостей в онлайн іграх, з метою отримання віртуальної валюти та/або цінних віртуальних предметів.

Атаки з використанням ботнет мереж – атаки даного типу базуються на виконанні однакового завдання на великій кількості пристроїв. Ботнет – це мережа комп'ютерів, заражена шкідливим програмним забезпеченням [27]. Зазвичай ботнет мережа організовується за допомогою поширення у мережі вірусного програмного забезпечення. Така мережа може складатися з різноманітних пристроїв (комп'ютери, телефони, побутові прилади, тощо). Найчастіше ботнет використовується для реалізації атаки типу DDoS для якої необхідно одночасно відправляти велику кількість запитів на один сервер [28]. З поширенням криптовалют ботнет мережі також почали використовуватися зловмисниками для майнінгу.

Атаки за допомогою машинного зору – метою даного виду атак є імітація роботи реального користувача. Зазвичай це використовується для збору інформації з веб-ресурсу або при необхідності подолати системи захисту, які базуються на тесті Т'юрингу. На сьогоднішній день, деякі «капчі» можна вирішити за допомогою ШІ. Таким чином був розроблений проект DeepCaptcha, який вирішує не лише прості капчі, де необхідно прочитати текст, а й графічні, які потребують вказати потрібні зображення.

Гібридні атаки – даний вид атаки комбінує розглянуті вище підходи для успішного проведення атаки зловмисником. Таким чином одні боти будуть збирати інформацію про жертву, інші тестувати зловмисні скрипти, а треті автоматизовано розгортати експлойти на пристрої жертви.

Розглянувши атаки спричинені ботами було виявлено, що на сьогоднішній день найчастіше зловмисники створюють власні ботнет мережі за допомогою вірусного програмного забезпечення. Після отримання достатньої кількості заражених пристроїв, подальші дії залежать від цілей зловмисника, але найчастіше виконуються саме атака типу DDoS.

У свою чергу технології парсингу постійно удосконалюються. Таким чином сучасні парсери біли «навчені» видавати себе за реального користувача, що значно ускладнює методи захисту від них.

Методи захисту від ботів

Методи захисту будь-якої системи ґрунтуються на прийнятій політики безпеці, яка забезпечує необхідні властивості системи захисту [29]. Особливості ботів полягають у тому, що вони можуть бути дозволеним або шкідливими. Як приклад дозволеного або корисного бота можна навести Googlebot, який автоматично

виявляє та сканує сайти, переходячи за посиланнями від сторінки до сторінки [30]. Однак при неправильному налаштуванні цього бота може сповільнюватися робота сервера, що викликає певний дискомфорт для користувача. На сьогоднішній день є лише один метод захисту від корисних ботів – використання загальноприйнятого серед розробників стандарту [31] виключень для ботів. Даний стандарт реалізовується через додавання файлу robots.txt до кореневого каталогу веб-ресурсу [32]. Головною проблемою цього стандарту є те, що його дотримання залежить лише від розробника бота. Це призводить до того, що лише невелика кількість корисних ботів дотримуються вимог стандарту.

Що стосується захисту веб-ресурсів від шкідливих ботів, зазвичай використовуються чорні списки, сигнатурне блокування, WAF та технологія Captcha.

Чорний список – це список IP-адрес та доменів, які були заблоковані через підозру в розсилці спаму. Спам – це небажані повідомлення у будь-якій формі, які надсилаються у великій кількості з використанням спеціальних ботів [33]. Найчастіше спам відправляється у формі комерційних електронних листів, надісланих на велику кількість адрес. Мета чорних списків – знизити відсоток надсилання небажаних розсилок користувачам [34]. Виділяють два типи чорних списків: на основі доменного імені відправника та на основі IP-адреси відправника.

Якщо домен або IP-адреса знаходиться в чорному списку, розсилки не будуть доставлені взагалі або пройдуть через додаткові спам-фільтри. Чорні списки використовують різні алгоритми ідентифікації спаму. На сьогоднішній день існує декілька відкритих чорних списків, а саме [35-37]:

– BRBL – безкоштовний чорний список IP-адрес, який створюється на основі заявок від користувачів;

– DNSBL – список IP-адрес та доменів, яких викрили в розсилці спаму;

– multiRBL – безкоштовний чорний список DNS, який посилається на інші чорні списки IPV4, IPV6 або домену;

– spamcop – список IP-адрес, які були відзначені користувачами як відправники спаму;

– spamhaus – список, який складається з декількох чорних списків DNS, у зв'язці з якими він відстежує та виявляє джерела небажаних розсилок, забезпечує захист від спаму;

– SURBL – списки сайтів, на які вели посилання в небажаних листах.

Сигнатурне виявлення шкідливих роботів – це спосіб виявлення і блокування роботів, який базується на аналізі їх характерних ознак, таких як шаблони HTTP-запитів, IP-адреси, заголовки, відбитки пристроїв [36]. Сигнатура бота – це його ідентифікатор, який дозволяє відрізнити його від легітимного трафіку. Сигнатурне виявлення шкідливих ботів може бути ефективним способом захисту від відомих і поширених ботів, але має свої недоліки.

Наприклад, сигнатурне виявлення шкідливих ботів неспроможна впоратися з новими чи невідомими ботами, які можуть змінювати свої сигнатури чи імітувати поведінка реальних користувачів. Також сигнатурне виявлення шкідливих роботів може призводити до помилкових спрацьовувань, якщо сигнатури роботів збігаються з сигнатурами легітимних відвідувачів. Тому такий спосіб виявлення шкідливих ботів рекомендується використовувати у поєднанні з іншими методами. Прикладом сигнатури бота може бути особливі поля у HTTP заколовку або відкритий текст, який складається з основних параметрів системи (розробник, дата створення, дата останнього оновлення, тощо) або хеш сумою цих даних.

WAF (Web Application Firewall) – це мережевий екран для фільтрації трафіку, який працює на прикладному рівні та захищає веб-додатки методом аналізу трафіку HTTP/HTTPS та семантики XML/SOAP [39]. WAF може встановлюватися на фізичний або віртуальний сервер та виявляє найрізноманітніші види атак. За принципом роботи, WAF працює як проксі-сервер. WAF може працювати з кластеризацією та акселерацією програм [40]. WAF може вбудовуватися в мережу як система моніторингу або шлюзу. Працює WAF за наступними моделями безпеки[39]:

– забороняючий режим. Система працює за заданим списком, який забороняє прийом конкретної інформації, прописаної в налаштуваннях. Захищає веб-застосунки на прикладному рівні (аналог IPS) та вміє оцінювати потенційні загрози детальніше і частіше застосовується для забезпечення захисту від «популярних» і специфічних типів атак. Аналізує вразливість конкретних веб-застосунків;

– дозвільний режим. Протилежність забороняючому режиму, тобто дозволяє приймати конкретну інформацію, яка була заздалегідь вказана в налаштуваннях. Дозволяє отримати максимальну захист.

CAPTCHA – це реалізація тесту Т'юринга, яка має на меті визначення, ким є користувач системи: людиною чи комп'ютером. Перша версія реалізації була розроблена на початку XXI століття інженерами з Університету Карнегі Меллона, США. Команда під керівництвом Луїса фон Ана (Luis von Ahn) шукала спосіб фільтрації реєстрацій на сайтах, яка виконуються автоматичними програмами та спам-ботами [41]. Команда розробила систему, яка показує користувачеві сильно спотворений текст, який неможливо розпізнати програмними алгоритмами. Роботу з ресурсом можна було продовжувати тільки після того, як слово, яке відображається, було коректно введено в поле тексту.

Рішення вийшло настільки вдалим, що стало активно застосовуватися у всьому світі. Однак це швидко призвело до появи нового виду заробітку в інтернеті - вирішення задач CAPTCHA. Спамери почали платити людям за введення «контрольної фрази». Такий дохід став популярним у бідних країнах, де можливість отримати мінімальні гроші за тисячі рішень CAPTCHA є досить привабливою [42].

Метод протидії роботі парсерів

Виходячи з аналізу історії парсерів та існуючих методів захисту було виявлено, що на даний момент не існує спеціалізованого методу захисту веб-додатку від даного виду шкідливих ботів. З цього виходить, що є необхідність розробки власного засобу протидії ботам даного виду. Для цього необхідно

проаналізувати особливості роботи парсингових систем. Після цього розробити алгоритм роботи системи захисту та розробити засіб, який буде вбудовуватися в існуючі веб-ресурси. Головною частиною будь-яких парсингових систем є набір правил, згідно яких система збирає дані. У цих правилах зазвичай вказане посилання на сторінку, з якої необхідно отримати інформацію та «шляхи» до елементів HTML сторінки, у яких зберігається необхідна інформація (текст, посилання або зображення).

На сьогоднішній день існують різні протоколи запису «шляхів» до елементів, серед них виділяють: XQuerry, CSSPath, RegEx та Xpath [43]. Серед представлених методів, шкідливі парсери найчастіше використовують систему запису Xpath. XPath — мова запитів до елементів документа XML. Ця мова розроблена для організації доступу до частин документа XML у файлах трансформації XSLT та є стандартом консорціуму W3C. XPath покликаний реалізувати навігацію по DOM у XML. У XPath використовується компактний синтаксис, відмінний від XML [44].

Кожна веб-сторінка обов'язково складається з структури у вигляді HTML документу та набору стилів, написаних на мові CSS. В процесі експлуатації та технічної підтримки розробники зазвичай не змінюють назви класів CSS та інші атрибути сторінки з метою стабільності роботи додатку. Цим і користуються зловмисники, які використовують парсингові системи для автоматизованого збирання даних. Так як атрибути HTML документів статичні, то і правила Xpath не потребують оновлення кожного дня.

Виходячи з цього необхідно створити умови, коли створення правил не буде мати сенсу. У даній роботі представлена система протидії, яка автоматично генерує випадкові імена атрибутів сторінки під час кожного запиту. Це призводить до того, що будь-які статичні системи парсингу перестануть працювати, так як після перезавантаження сторінки існуючі правила вже не будуть працювати. Алгоритм роботи засобу представлений на рис. 1. Згідно ньому, при отриманні запиту GET від користувача, сервер запускає створення сторінки. Під час цього процесу система генерує випадкові назви для HTML-атрибутів. Після чого сторінка відправляється користувачу.

Висновки

Під час аналізу атак на комп'ютерні системи, для проведення яких використовувалися автоматизовані системи було виявлено, що у більшості випадків ботів використовують для проведення атаки типу DDoS. Для цього використовуються ботнет мережі. Для ефективного створення такої мережі, зловмисники створюють програмне забезпечення, яке заражає не лише домашні пристрої, а й елементи IoT. Це зумовлено тим, що виробники таких пристроїв доволі рідко створюють оновлення системи, що призводить до появи великої кількості вразливостей. Також було виявлено, що атаки за допомогою краулерів майже неможливо виявити, оскільки їх важко відрізнити від справжніх користувачів. Тому всі випадки атак за допомогою таких систем ставали відомі лише завдяки появі вкраденої інформації на спеціалізованих ресурсах для продажу вкраденої інформації.

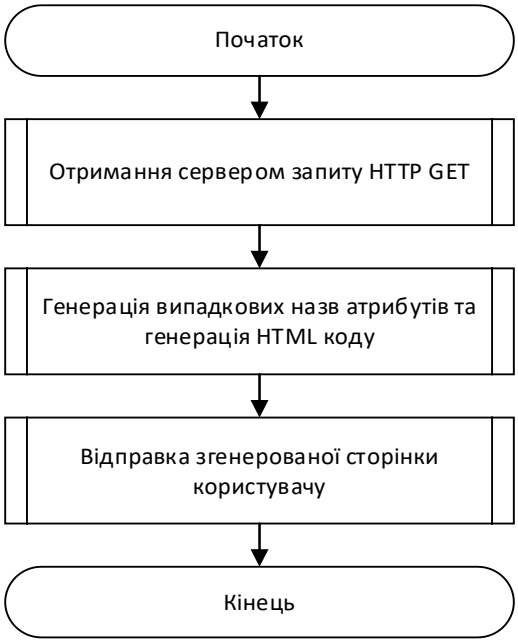

Рис. 1. Алгоритм роботи засобу протидії парсерам

Під час аналізу методів захисту від автоматизованих систем було виявлено, що на сьогоднішній день, основною проблемою під час розробки методів захисту є розмежування реальних користувачів від ботів. В залежності від складності алгоритму роботи бота, деякі системи захисту можуть бути або неефективними, або навіть заважати справжнім користувачам отримувати інформацію на ресурсі. У свою чергу, на сьогоднішній день відсутні засоби, які спеціалізуються саме на захисту від парсингових систем.

Запропоновані методи спеціалізуються на захисті від парсингових систем, та створює мінімальне додаткове навантаження на серверне обладнання, що не заважає роботі звичайним користувачам. Даний метод буде корисним власникам великих ресурсів, на яких головним є саме інформація.

Такими ресурсами можуть бути сайти новин, інтернет магазини або навіть сайти соціальних мереж. Це зумовлено тим, що за статистикою дані сайти найчастіше підвержені атакам зі сторони парсерів, а тому представлений метод зможе не лише ефективніше використовувати обчислювальну потужність серверів, а й вирішити проблеми авторських прав, коли один ресурс заробляє на інформації, розташованій на іншому.

СПИСОК ЛІТЕРАТУРИ

1. Pevnev V., Frolov A., Tsuranov M., and Zemlyanko H. Ensuring data integrity in infocommunication systems. International Journal of Computer Science, 21(2), 2022. pp.228–233. doi.org/10.47839/ijc.21.2.2591;
2. Статистика ботів. Begibot. URL: https://www.begindot.com/ua/
3. Imperva Bad Bot Report. Imperva. URL – https://www.imperva.com/resources/reports/2023-Imperva-Bad-Bot-Report.pdf

4. A. Serkov, V. Tkachenko, V. Kharchenko, V. Pevnev, K. Trubchaninova, N. Doukas, “Method of increasing security of spatial intelligence in the industrial internet of things systems,” Proceedings of the 24th Int. Conference on Circuits, Systems, Communications and Computers, CSCC’2020, 2020, pp. 283–289. https://doi.org/10.1109/CSCC49995.2020.00058;
5. Serkov, A., Tkachenko, V., Kharchenko, V., Pevnev, V. Method to Enhance the Bandwidth and Noise Immunity of IIoT When Exposed to Natural and Intentional Electromagnetic Interference. 2020 IEEE Int. Conf. on Problems of Inf. Science and Technology (PIC S&T). Kharkiv, 2020. p.527–532. doi: https:// doi.org/10.1109/picst51311.2020.94679295
6. Instagram тестує нові попередження. Unian. URL: https://www.unian.net/techno/iskusstvennyy-intellekt-ne-proydet-instagram-testiruet-novye-preduprezhdeniya-12348897.html
7. Чат-бот. Sendpulse. URL: https://sendpulse.ua/ua/support/glossary/chatbot
8. Що таке веб-краулер? Brightdata. URL: https://ua-brightdata.com/blog/web-data-ru/what-is-a-web-crawler
9. ALGOL. Ain. URL: https://ain.ua/ua/2021/09/24/5-mertvix-jazikov-programmirovanija/
10. Пошуковий індекс. Roistat. URL: http://surl.li/qrnqh.
11. Чат-бот «Еліза» з 1960-х років пройшов тест Т'юрінга краще, ніж ChatGPT. Technovery. URL: https://technovery.com/chat-bot-eliza-iz-1960-h-godov-proshel-test-tyuringa-luchshe-chem-chatgpt/
12. Beagle. DBpedia. URL – https://dbpedia.org/page/Bagle_(computer_worm)
13. Akbot. DataProtection. URL: https://vms.dataprotection.com.ua/virus/?i=95482
14. Виявлено ботнет Win32/Georbot, який використовує для оновлення сайт уряду Грузії. ESET. URL: https://www.eset.com/ua-ru/about/newsroom/press-releases/malware/obnaruzhen-win32-georbot-napadenie-ru/
15. Izz ad-Din al-Qassam Cyber Fighters. Radware. URL – https://www.radware.com/security/ddos-knowledge-center/ddospedia/izz-ad-din-al-qassam-cyber-fighters/
16. Ботнет Mirai. Enigmasoftware. URL: https://www.enigmasoftware.com/ua/mirai-botnet-udaleniye/
17. 8 найбільших DDoS-атак в історії. Timeweb. URL: http://surl.li/qrnql.
18. Securing Broncos Country. Checkpoint. URL – https://www.checkpoint.com/security-in-action/
19. Cybersecurity news from Hong Kong. Portswigger. URL – https://portswigger.net/daily-swig/hong-kong
20. Malicious attacks on the web and crawling of information data by Python technology. URL https://www.researchgate.net/publication/351772882_Malicious_attacks_on_the_web_and_crawling_of_information_data_by_Python_technology
21. Хакери знову напали на американські банки. Finance.Bigmir. URL: https://finance.bigmir.net/news/2824135
22. Protection from even the most severe DDoS attacks. Stormwall. URL – https://stormwall.network/
23. Актуальні кіберзагрози: IV квартал 2023 року. Fortiguard. URL: https://www.ptsecurity.com/ru-ru/research/analytics/cybersecurity-threatscape-2022-q4/
24. Anti-Botnet Services. Fortiguard. URL – https://www.fortiguard.com/services/botnet
25. EnemyBot. Enigmasoftware. https://www.enigmasoftware.com/ua/enemybot-udaleniye/
26. Що таке ботнет? ESET. URL: https://www.eset.com/ua-ru/support/information/entsiklopediya-ugroz/zashchita-ot-botnetov/
27. Розвиток ботнетів і DDoS-атак. IITD. URL: https://iitd.com.ua/ua/news/rozvitok-botnetiv-i-ddos-atak/
28. Crypto Mining Bot. Netacea. URL – https://netacea.com/glossary/crypto-mining-bot/
29. Pevnev V., Tsuranov M., Zemlianko H., Amelina O. Conceptual Model of Information Security, Integrated Computer Technologies in Mechanical Engineering, 2020, Vol. No 188, pp. 158–168. DOI: 10.1007/978-3-030-66717-7_14;
30. Загальні відомості про наші пошукові роботи та інструменти для збору даних. Google Developers. URL: https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers?hl
31. Uniform Resource Identifier. RFC. URL – https://www.rfc-editor.org/rfc/rfc3986
32. Що таке robots.txt і навіщо взагалі потрібний індексний файл. Netpeak. URL: https://netpeak.net/ua/blog/
33. Як перевірити IP-адреси сервера та домену в спам-базахю. Unisender. URL: http://surl.li/qrnrh.
34. Перевірка IP-адрес у спам-листах. Ukraine.com. URL: https://www.ukraine.com.ua/info/tools/rbl
35. Email Blacklist Перевірка. BRBL. URL: https://ipcalc.co/rbl/
36. What is a DNSBL? DNSBL. URL – https://www.dnsbl.info/
37. Spamhaus. Spamhaus Project. URL – https://www.spamhaus.org
38. Захист від ботів з PT Application Firewall. Slideshare. URL: https://www.slideshare.net/VsevolodPetrov/pt-application-firewall
39. WAF. ITglobal. URL: https://itglobal.com/ru-ru/company/glossary/waf/
40. Web Application Firewall. Omnilink. URL: https://omnilink.ua/web-application-firewall/
41. Що таке CAPTCHA? Google. URL: https://support.google.com/a/answer/1217728?hl=ru
42. BestCaptchaSolver. Bestcaptchasolver. URL: https://bestcaptchasolver.com/
43. Regular Expressions in XQuery: A Rephrased Perspective. CopyProgramming. URL – https://copyprogramming.com/howto/xquery-regular-expressions
44. Що таке XPath? Функції та синтаксис. HighLoad. URL: https://highload.today/xpath-xml

Received (Надійшла) 13.02.2024
Accepted for publication (Прийнята до друку) 10.04.2024

Methods of preventing automated collection of information from web pages

O. Yudin, M. Shypunov

Abstract. The work is devoted to the analysis of methods of counteracting the automated scanning of web sites. The purpose of the work is to analyze the algorithms and features of parsing systems and, based on the data obtained, to build a tool system that will specialize in detecting and countering attacks using parsing systems. The research method is the analysis of methods of countering parsing systems. The paper considered the history of the emergence of automated systems, their classification, features of work and methods of countermeasures. The proposed methods specialize in protection against parsing systems, and create a minimal additional load on server equipment, which does not interfere with the work of ordinary users. These methods will be useful to the owners of large resources, on which information is the main thing.

Keywords: parsing software, anti-bot methods, parsers, crawlers.